DOCUMENT RESUME

ED 433 362 TM 030 051

AUTHOR	Chiu, Christopher W. T.; Pearson, P. David
TITLE	Synthesizing the Effects of Test Accommodations for Special Education and Limited English Proficient Students.
SPONS AGENCY	American Institutes for Research in the Behavioral Sciences, Palo Alto, CA.
PUB DATE	1999-06-00
NOTE	51p.; Paper presented at the National Conference on Large Scale Assessment (Snowbird, UT, June 13-16, 1999).
PUB TYPE	Information Analyses (070) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Academic Achievement; Achievement Gains; Achievement Tests; *Disabilities; Elementary Secondary Education; *Limited English Speaking; Meta Analysis; Research Reports; *Special Education; Test Results; *Testing Problems; Timed Tests
IDENTIFIERS	*Testing Accommodations (Disabilities)

ABSTRACT

Test accommodations for special education (SP) and limited English proficient (LEP) students have attracted much attention recently because proper accommodations promote inclusion and allow students to perform optimally. A meta-analysis of 30 research studies found empirical evidence supporting the position that, with appropriate accommodations, SP and LEP students can increase their scores on standardized achievement tests. Compared to conditions of no accommodation, students increased their scores by an average of 0.16 standard deviation. Relative to general education students, accommodated SP and LEP students demonstrated an average accommodation advantage of 0.10 standard deviation. Interpretations of these average effects require careful analyses because of the variety of accommodations, the specific status of the students, and the varying implementations of the accommodations. Providing additional time or unlimited time is the most frequently investigated accommodation. Other accommodations investigated were assistive devices, presentation formats, response formats, test settings, radical accommodations, and combinations of accommodations. Age did not seem to be a factor; elementary and postsecondary students benefited from accommodations. Narrative descriptions are given of the situations in which positive and negative effects of accommodation appear to emerge. An appendix lists and summarizes the studies analyzed. (Contains 63 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Running Head: Effects of Test Accommodations

ED 433 362

Synthesizing the Effects of Test Accommodations for Special Education and Limited English Proficient Students

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Christopher W.T. Chiu

Christopher W. T. Chiu

P. David Pearson

Department of Counseling, Educational Psychology, and Special Education

Michigan State University

Note

Paper to be presented to the 1999 National Conference on Large Scale Assessment (June 13 - 16, Snowbird, UT) organized by the Council of Chief State School Officers (CCSSO). Christopher W.T. Chiu specializes in Measurement and Quantitative Methods; P. David Pearson, John A. Hannah Professor at Michigan State University. This project was funded by the American Institute for Research (AIR). Correspondence concerning this article should be addressed to Christopher W.T. Chiu, 5C01 Forbes Quad, University of Pittsburgh, Pittsburgh, PA 15260. Electronic mail may be sent via Internet to chiuwing@pilot.msu.edu. Telephone: (517) 337-7486. The authors thank Lorrie Shepard, Gerald Tindal, the National Center on Educational Outcomes, and many other researchers who have contributed research studies to this paper. The authors also thank those who provided suggestions and consultations to our meta-analysis including the SynRG research group for meta-analysis at Michigan State University. Constructive comments by Larry Hedges are gratefully acknowledged.

TM030051

BEST COPY AVAILABLE

ABSTRACT

Test accommodations for special education (SP) and limited English proficient (LEP) students have attracted much attention recently because proper accommodations promote inclusion and allow students to perform optimally. Using a procedure outlined by Hedges and Olkin (1985), we conducted a meta-analysis of 30 research studies and found empirical evidence supporting the position that, with appropriate accommodations, SP and LEP students can increase their scores on standardized achievement tests. We found that compared to conditions of no accommodation, these students increased their scores by an average of 0.16 standard deviation. Relative to general education students, accommodated SP and LEP students demonstrated an average accommodation advantage of 0.10 standard deviation. Interpretations of these average effects require extremely careful analyses because a wide variety of accommodations exist, the statuses of students are specific, and the implementations of accommodations vary in nature and quality. Providing extended time or unlimited time to students with learning disabilities is by far the most frequently investigated accommodation; students with learning disabilities are the most commonly investigated target population. Other accommodations investigated were *assistive devices*, *presentation formats*, *response formats*, *setting of tests*, *radical accommodations* and *combinations of accommodations*. Age did not seem to be a factor in the analysis; both K-6 and postsecondary students tended to benefit from accommodations. In order to provide guidance for those charged with the responsibility of implementing accommodations, we close our synthesis with narrative descriptions of the characteristics and situations in which both positive and negative effects tend to emerge.

INTRODUCTION AND PROBLEM FORMULATIONS

In the United States, special education students and limited English proficient students (LEP) constitute a large portion of the student population. According to a recent national report (Olson & Goldstein, 1997), more than four million students ages 6 - 21 have at least one disability (p. 13) and more than two million students from kindergarten to postsecondary levels have limited English proficiency (p. 39). According to the New York Times (Tamar, 1998), more than five percent of the nation's students now have diagnosed learning disabilities.

In addition to the needs of special education and LEP students, federal laws and parents involvement are major sources advocating inclusion and accommodations. Some states such as Maryland and Virginia started to grant special education students diplomas when they graduate from high schools (Perlstein, 1999). Other states including Louisiana mandate that special education and / or LEP students participate in their statewide schools accountability programs (MacGlashan, 1999). Parents concerned that their children receiving special education should be tested fairly when examinations are used for making promotion and any other high-stake decisions (MacGlashan, 1999). Numerous students, who have diagnosed learning disabilities, in elementary schools, colleges and graduate schools, and professional schools receive accommodation when taking tests (Tamar, 1998). The Americans with Disabilities Act (ADA) requires that all students with disabilities to be provided the same educational opportunity with those without disabilities. As far as accountability is concerned, the Individuals with Disabilities Education Act Amendments of 1997 (P.L. 105-17) require states to report on performance of students with disabilities. This requirement motivates states to actively involve in improving academic and non-academic outcomes of the special education students (Ysseldyke, Thurlow, Kozleski & Reschly, 1998). It also raises an urgent issue regarding the needs of accommodating special education and LEP students when they take a test.

Inclusion has not been widely implemented until recent years because of several challenges (Heubert & Hauser, 1998; Olson & Goldstein, 1997). Most of these challenges are related to the controversial issues of the appropriateness, impact, and validity of offering accommodations to special education and LEP students.

Many believe that the logic behind accommodations is or ought to be the removal of variances irrelevant to the constructs being measured, so that the true competence of the target students can be measured accurately (e.g., Philips, 1994; McDonnell, McLaughlin, and Morison, 1997). That is, with appropriate accommodations, a student disability or LEP status, if unrelated to the constructs being measured, will no longer be a source hindering their true demonstration of competence. Without accommodations, target students may score lower than they should. Test accommodations are generally thought of as a corrective lens for such potential score distortions (McDonnell, McLaughlin, and Morison, 1997; Philips, 1994). However, providing test accommodations remains a debatable topic because of the scarcity of research evidence to support the effects, validity, and psychometric properties of the various categories of accommodations (Heubert & Hauser, 1998; McDonnell, McLaughlin, and Morison, 1997; Olson & Goldstein, 1997; and Thurlow, Ysseldyke, & Silverstein, 1993). In addition, the operational definitions of disabilities (e.g., learning disabilities) and English proficiency level can be elusive (e.g., Olson & Goldstein, 1997; Swanson, 1991).

Among the aforementioned challenging issues, the examination of the effects of test accommodation is an urgent issue that attracts much attention at national conferences in large-scale assessments (e.g., the 1997, 1998, and 1999 Council of Chief State School Officers Annual Meetings on Large Scale Assessment). Calls for reports on the effects of test accommodations have proliferated (Ysseldyke, Thurlow, Kozleski & Reschly, 1998; Thurlow, Ysseldyke & Olsen, 1999). The questions that plague policymakers and researchers include:

- What constitutes an effective accommodation?
- To what extent can an accommodation improve the scores of target students?
- Should accommodation be provided to only target students or to all students?

In the current study, we address these questions by using a meta-analysis (Hedges and Olkins, 1985) to synthesize the findings of studies that have evaluated the magnitude of the score improvement affected by a number of the most popular accommodations examined in the 30 empirical studies that emerged from our search of the professional literature (which began with a set of about 200 potential candidates). We began with the belief that, as a field, we had accumulated enough empirical research on accommodations to make meta-analysis useful; that is, we believed that the database was sufficiently large to take advantage of what meta-analysis offers—the capacity to examine convergence across studies in as systematic and disinterested a manner as possible. Also, by converting score scales used in different studies to a common metric (effect sizes), meta-analytic techniques allow one to determine common characteristics contributing desirable or undesirable effects to a particular study. In short, we believed that we were at a point at which we could begin to provide an empirically based answer to the question, What makes an accommodation effective or ineffective?

METHODS

Selection Criteria for Empirical Studies

We first developed a framework for including empirical studies for analysis by adopting the broad range of accommodations that are currently in use by states and school districts (Bechard, 1997; Neuburger, 1997; Olson & Goldstein, 1997; Porter, 1997; Thurlow, 1997; Thurlow, Ysseldyke, & Silverstein, 1993; Trent, 1997). Then we settled on including groups of special education and limited English proficient students in the research synthesis. We started with a four-category taxonomy on the nature of disabilities (Willingham, Ragosta, Bennett, et al., 1988), including visual impairment, hearing impairment, physical disabilities, and learning disabilities. Soon after a preliminary coding, we expanded

the taxonomy to include an even more diverse construal of disabilities, adding hyperactive students and those who, while academically at risk, do not have a formal status of disabilities (e.g., garden-variety reading disabled students).

Regarding the accommodations provided to LEP students, we adopted the taxonomy used by various states throughout the nation (Olson and Goldstein, 1997). We then discussed the coding scheme for high inference variables (i.e., types of accommodations, taxonomy of disabilities and LEP status, quality of research designs) and reached an 87% agreement based on 15 randomly selected studies. Disagreements were discussed and resolved, and the resolution was subsequently applied to the entire coding process. Having realized that some factors like cultural issues (Garcia and Pearson, 1994) and lack of prior knowledge (unfamiliarity, e.g., historical background in a reading proficiency test) could become barriers for LEP students in their test performance, we broadened the scope of the selection criteria for accommodations. In order to determine if the accommodations under investigation "matched" the needs of the target students, we checked to ensure that the included research studies had explicitly described the nature of the target students and had provided narrative descriptions of the accommodations used. We followed the guidelines suggested by Swanson (1991) and Olson and Goldstein (1997) to confirm that the research studies we included investigated target students with known disabilities or LEP statuses. Examples of the various accommodations provided in the empirical studies are provided as follows:

- ***Timing of Test***: provided extended-time or unlimited time.
- ***Radical Accommodations***: 1) provided students with background knowledge (particularly culture-specific information needed for full understanding) of a reading test, such as the story's title, the author, and the year of publication; 2) reduced cultural loads on reading passages for LEP students in a reading test (i.e., choose a universal theme rather than a theme specific to the U.S.); 3) provided hypermedia reading aids (added speech synthesis, an on-line glossary, links between questions and text, highlighting of main ideas, and supplementary explanations that summarized important ideas) to students with reading difficulties on a biology test.

- ***Presentation Formats***: 1) presented a mathematics test using a computer for students with learning disabilities (LD); 2) used video presentations for math word problems;
- ***Assistive Devices***: 1) provided graphical organizers (to organize outlines) to LD students on social science tests; 2) glossaries for LEP students on mathematics tests; 3) lenses for visually impaired students.
- ***Combinations of Accommodations***: 1) supplied large print and extended time to visually impaired students; 2) provided combinations of a separate location, extra time, a reader, a transcriber, an interpreter, rest periods, or special equipment; 3) audiotaped the test and provided extended time; 4) allowed a glossary and extra time for LEP students.
- ***Response Formats***: 1) allowed students with cerebral palsy to respond to a reading test by touching a single-switch device.
- ***Setting of Test***: 1) administered an arithmetic exam to hyperactive students under settings in which 10 minutes of music was played, or 10 minutes of background speech was played when they were taking the test.

At our first cut, we had 111 research studies based on the relevance of their titles and abstracts. We then eliminated studies with serious methodological flaws, such as a lack of random assignment or the use of research designs that threatened external validity (Bangert-Drowns, 1993) to ensure the quality and the generalizability of the research synthesis. Case studies and studies that did not provide sufficient information for effect-size computation (e.g., standardized mean differences) were also excluded from the meta-analysis. We, however, kept these studies in our overall database so that we could archive and summarize them; we then used the insights they provided to inform our overall discussion about the nature of accommodations.

Review of Literature

Our search uncovered 30 empirical studies directly relevant to the testing of accommodations and met the selection criteria discussed earlier. Ten studies were excluded from the meta-analysis because of any or all of the following reasons:

1. The accommodations provided altered the underlying construct(s) being measured (e.g., sample accommodations include the simplification of sentence structure in listening tests and the use of corrective feedback for oral reading tests);
2. Outcome measures tended to be unreliable (e.g., use of classroom grades given by different teachers within a short duration);
3. Studies had serious methodological flaws (e.g., use of post hoc analysis to define groups that received accommodations for language testing).

We found the empirical studies in a variety of sources, including journals, dissertations, national conferences, databases, research reports, and personal communications. Regarding national conferences, we searched through programs of recent annual meetings of: (1) the American Educational Research Association, (2) the National Council on Measurement in Education, and (3) the National Conference on Large Scale Assessment. We also exhausted the dissertation abstracts database, an invaluable source for empirical work. Whenever dissertation abstracts reported relevant information, we ordered and read them in entirety. We browsed through the listings of technical reports published by the American College Testing (ACT) program and the Educational Testing Service (ETS); relevant reports were ordered and read. We also searched through journal articles published in the past 12 years (1986-1998). See Table 1 for the major sources from which we retrieved the empirical studies.

Table 1: Sources of the Empirical Studies

	A	B	C
1	Source	Titles	Coverage
2	Journals	Journal of Special Education	1986 – 1997
3		Journal of Learning Disabilities	1987 – 1998
4		TESOL Quarterly	1987 – 1997
5		Journal of Educational Measurement	1990 – 1998
6		Reading Research Quarterly	1990 - 1998
7		Journal for Research in Mathematics Education	1991 – 1998
8		Supplement to the JRME: Annual Listing of Research on Mathematics Education	1990 – 1994
9		Research on Journal of Reading Behavior	1990 - 1998
10		Review of Educational Research	1990 - 1998
11	Reports	National Center on Educational Outcomes (NCEO)	Web Site on the Internet
12		CRESST	Web Site on the Internet
13		American College Testing (ACT)	Browsed the publication list containing reports in and prior to 1997
14		Educational Testing Service (ETS)	Obtained research reports on testing of handicapped students in the 1980s
15	Review Studies	National Center for Educational Statistics (NCES)	Contacted for relevant reports
16		Thurlow, Ysseldyke, and Silverstein (1993)	Reference list searched
17		Olson and Goldstein (1997)	Reference list searched
18			
19	Conferences	CCSSO	1997 – 98
20		AERA/NCME	1997 – 98
21	Dissertations	Dissertation Abstracts	Keywords used: accommodations, assessments, braille, disabilities, extra time, formats, learning disabilities, presentation formats, second language, test modifications, testing, timed, untimed

Empirical Studies Found

We conducted a meta-analysis on the 30 studies with acceptable research designs and sufficient statistical information (i.e., means, standard deviations, t-, or F- statistics) to merit inclusion in our sample. The majority of student participants in the 30 studies was in the K-6 level (n=8) and in the postsecondary level (n=16). Nine studies focused on middle school and high school students and four studies selected students from across a wide range of levels. All the studies examined one or more academic achievement(s) as outcome measures, including mathematics, reading, writing, listening, social studies, and science. Reading and mathematics comprised the major subject areas among the studies.

BEST COPY AVAILABLE

Accommodation by Disabilities

Students with learning disabilities (LD) was the most frequently studied group in the sample of studies (n=19). LEP / English as a Second Language (ESL) / Cultural subgroups ranked second (n=7). As reported in the second to last column of Table 2, *timing of test* seems to be a universal accommodation that could be and was provided to a variety of special education students, including LD students, LEP/ESL/cultural subgroups, hyperactive students, and physically disabled students. *Presentation format* and *combinations of accommodations* were two other types of practical accommodations that could be and were provided to various special education students. *Assistive devices* and *setting of test*, were, however, accommodations given to meet the needs of students with particular disabilities; for example, being tested in a separate room (*setting of test*) was provided only to hyperactive students.

Table 2: Frequency of Accommodation Used by Special Education Status

	A	B	C	D	E	F	G	H	I
1		Assistive Devices	Combinations of Accommodations	Presentation Format	Radical Accommodations	Response Format	Setting of Test	Timing of Test	Total
2	ESL / LEP / Cultural Subgroups		1	1	2			3	7
3	Students with Garden Variety Disabilities			1				1	2
4	Hyperactive Students						1		1
5	LD Students	1	4	2	2			10	19
6	Multiple Groups of Special Education Students		2					2	4
7	Physically Handicapped Students							1	1
8	Students with No Formal Disability Status			1		1			2
9	Visually Impaired Students	3						1	4
10	Total	4	7	5	4	1	1	18	40

Research Designs

Among the 30 studies with sufficient statistics for meta-analysis, the most commonly used research design for test accommodations was the *repeated measure with a comparison group* (n = 16). In this design, the target population students were tested in both standard and accommodated conditions. Regular education students with similar background, grade levels, and IQ were selected as a comparison group, and they were also tested in both conditions. The *repeated measure without a comparison group* (n

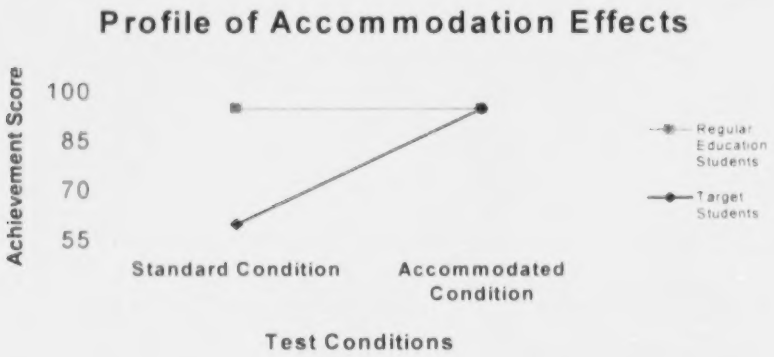
 BEST COPY AVAILABLE

= 7) is another commonly employed design, especially for accommodations that are illogical to offer to non-target education students (e.g., administering a Braille version of a reading test to sighted students). The *equivalent group design* (n = 7) consisted of studies in which students from the target population were randomly divided into two groups, with one group tested under the standard condition and the other group tested in the accommodated condition.

Effect Size Computations

A major concern regarding whether a given accommodation is acceptable turns on the extent to which test scores differ between the standard and accommodated conditions for special education as well as for regular education students (Mehrens, 1994 and Phillips, 1992). Researchers (e.g., Phillips, 1994) argue that the function of an acceptable accommodation is to remove the irrelevant barriers affecting the skills being measured. That argument implies that if an accommodation works as intended, the test scores of special education students should improve and the test scores of regular education students should not change significantly. Consistent with Phillips (1994), Shepard, Taylor, Betebenner, and Weston (in press) and Tindal, Heath, Hollenbeck, et al. (1998) point out that an "effective" accommodation should have a profile exhibiting an interaction effect between test conditions and the statuses of the students. That is, target students should improve their scores in an accommodated condition compared to their scores in a standard condition, while non-target students' scores should remain unchanged or change only slightly. The standardized mean change effect size (Becker, 1987) is well suited to portray such a profile because it can be used to investigate the extent to which special education students can achieve scores higher (or lower) than regular students prior to and after an accommodation is provided. Another advantage of the standardized mean change effect size is that it converts outcome measures developed on different scales (e.g., different reading tests) to a comparable metric (standard deviations units). Figure 1 depicts what we expect to occur when an accommodation exhibits an optimal effect:

Figure 1: A Hypothetical Example Showing Expected Accommodation Effects



In addition to the graphical presentation in Figure 1, we computed effect sizes for regression analyses, which allowed us to make predictions of the effects of a given accommodation. Equation 1 shows the computational formula for the standardized mean change effect size (what we called the $g_{relative\ effect}$), which is based on the difference of two effect sizes (g_{target} and $g_{regular\ ed}$).

$$g_{relative\ effect} = g_{target} - g_{regular\ ed} \quad (1)$$

$$g_{target} = \frac{\hat{\mu}_{target\ accommodated} - \hat{\mu}_{target\ standard}}{\hat{\sigma}_{pooled\ target}} \quad (2)$$

$$g_{regular\ ed} = \frac{\hat{\mu}_{regular\ ed\ accommodated} - \hat{\mu}_{regular\ ed\ standard}}{\hat{\sigma}_{pooled\ regular\ ed}} \quad (3)$$

As can be observed in Equation 1, the effect of test accommodation, $g_{relative\ effect}$, is defined by the difference between the standardized mean change effect size of the special education students and that of the general education students. A positive value indicates that special education students benefit more from a given accommodation than do general education students. However, the effect size $g_{relative\ effect}$ is insufficient to be used as the sole indicator for the effect of accommodation, because $g_{relative\ effect}$ can be

BEST COPY AVAILABLE

positive even if the special education students do not benefit from a given accommodation. For example, if the g_{target} effect was -1 and the $g_{regular\ ed\ effect}$ was -2, the $g_{relative\ effect}$ would be +1. To assist interpretation, we also report the effect size for the target students (g_{target}). Optimally, both $g_{relative\ effect}$ and g_{target} will be positive.

Variance Computations

Correction for bias — Because the sample size is generally small for studies in special education, correction for bias introduced by the small sample size is essential to estimate the population effect. We adopted Hedges and Olkin's (1985) formula to correct for the bias. The formula is summarized as follows: $c(m) = 1 - 3 / (4m - 1)$, where m is the sum of the sample sizes of the two groups. We used this formula to unbais the effect sizes (g_{target}, $g_{regular\ ed}$, and $g_{accommodated}$) that are based on a match group design. For effect sizes that are based on a repeated measure design (one in which the same sample is used in accommodated and standard conditions), we adjusted the Hedges and Olkin's formula by defining m as the average of the sample size of the accommodated and standard conditions.

Computing the variance for effect sizes — For the repeated measure design, we include in the variance formula the correlation between the accommodated and standard conditions: $\text{var}(d_i) = 2 \bullet (1 - \rho_{AS}) / m + d_i^2 / (2 \bullet m)$. Note that d_i is the effect size for primary study i; ρ_{AS} is the correlation of the scores between the accommodated and standard conditions; and m is the average of the sample sizes in the two conditions. For the equivalent group design, we did not need to include the correlation between the accommodated and standard conditions because the samples in the equivalent group design are independent. We used the following formula to compute the variance: $\text{var}(di^*) = (n_{i1} + n_{i2}) / (n_{i1} \bullet n_{i2}) + d_i^{2*} / 2 \bullet (n_{i1} + n_{i2})$. Notice that n_{i1} and n_{i2} are the sample sizes of the two different groups in the equivalent group design.

RESULTS

Overall Effects of Accommodations

As is reported in Table 3 on page 15 (see cell I 15) *timing of test* was the most frequently investigated accommodation. Almost half (47%) of the accommodations provided extended time or unlimited time. *Setting of test* (2%) and *response format* (2%) were the least frequently investigated accommodations. Four other frequently examined accommodations included *assistive device* (9%), *combination of accommodation* (11%), *presentation formats* (13%), and *radical accommodation* (17%). Table 4 (p.19) shows the breakdown by the different types within the target population. Comparing across the subgroups in the target population, the LD subgroup was most widely studied (61%) and the ESL / LEP subgroup was second (16%). The LD subgroup was not only more widely studied than the ESL / LEP subgroup, it was also more frequently studied with *timing of test* as an accommodation.

While we were primarily interested in whether there was an overall effect for test accommodations, we were mindful of the possibility that the effects of accommodation vary depending on the accommodation and the students. Hedges and Olkin (1985) have outlined a procedure to examine the relationship between the characteristics of studies and outcome measures. Their procedure, analogous to Analysis of Variance, computes an average effect size of the test accommodation effect, weighted by the inverse of the variances of the individual effect sizes. A test statistic called Q is used to determine the homogeneity of the variance among effect sizes within the group of studies under consideration. This statistic is called Q_W when applied to a particular category, such as LD students, time allocated for the test, or a particular age band of students. The Q_W has an approximate chi-square distribution with a degree of freedom k - 1 where k is the number of effect sizes. A significant Q_W for a particular category indicates that the variation within the category is sufficiently large such that it is not possible to consider the studies as "belonging" to the same category. This Q statistic was applied to the entire set of effect

the studies as "belonging" to the same category. This Q statistic was applied to the entire set of effect sizes in our sample (n = 31 for the relative accommodation effects and n = 47 for the accommodation effects for the target population) as well as to each relevant subcategory (e.g. presentation format, timing, learning disabled students, ESL students, K-6 students, and postsecondary students).

Table 3: Average Test Accommodation Effects: Categorical models for assessing various moderators on the effects of accommodations

	Accommodation Effects (Effect3: weighted average = 0.11, unweighted average = 0.48)						*Benefit to Special Education Students (Effect1: weighted average = 0.16, unweighted average = .58)*					
Variable and group	No d's	Qw	95% CI Lower	95% CI Upper	Weighted d's	Unweighted d's	No d's	Qw	95% CI Lower	95% CI Upper	Weighted d's	Unweighted d's
Accommodations												
1 Assistive device	2	11.98	0.23	0.38	**0.07**	0.73	4	42.34	0.2	0.51	**0.35**	0.71
2 Combinations of accommodation	1	N/A	0.72	0.38	**-0.17**	0.17	5	3.06	0	0.21	**0.1**	0.12
3 Presentation formats	4	3.68	0.00	0.23	**0.12**	0.12	7	116.35	0.18	0.06	**-0.12**	0.06
4 Radical accommodations	4	19.26	0.02	0.51	**0.24**	0.71	7	37.06	0.14	0.42	**0.28**	0.75
5 Response format	1	N/A	0.68	0.46	**-0.11**	0.11	1	N/A	0.04	0.38	**-0.02**	0.02
6 Setting of test	1	N/A	0.27	1.1	**0.68**	0.68	1	N/A	0.16	0.78	**0.47**	0.47
7 Timing of test	18	70.7	0.02	0.16	**0.07**	0.53	22	124.64	0.26	0.35	**0.31**	0.80
Special education students status												
1 ESL / LEP / cultural subgroups	5	6.52	0.07	0.27	**0.10**	0.03	10	10.19	0.15	0.26	**0.21**	0.18
2 Garden variety	2	0.14	0.12	0.18	**0.03**	0.00	2	3.65	0.07	0.09	**0.01**	0.02
3 Hyperactive	1	N/A	0.27	1.1	**0.68**	0.68	1	N/A	0.16	0.78	**0.47**	0.47
4 Learning disabilities	19	100.25	0.01	0.22	**0.11**	0.72	23	379.51	0.06	0.18	**0.12**	0.91
5 Multiple groups of LD students	0			N/A			4	2.1	0.02	0.2	**0.09**	0.08
6 Physically handicapped	1			N/A			1	N/A	0.35	0.93	**0.64**	0.64
7 No formal status	4	0.79	0.04	0.24	**0.10**	0.03	4	11.76	0.08	0.29	**0.19**	0.49
8 Visually impaired	0			N/A			2	12.02	0.41	0.83	**0.62**	0.48
Grade level												
1 K - 6	8	9.81	0	0.18	**0.09**	0.10	8	128.89	-0.13	0.02	**-0.07**	0.24
2 Middle school (7 - 8)	7	23.38	-0.13	0.23	**0.05**	0.35	7	38.48	0.02	0.27	**0.14**	0.65
3 High school (9 - 12)	0			N/A			2	1.19	0	0.44	**0.22**	0.30
4 Postsecondary (college and university)	15	72.24	0.02	0.24	**0.13**	0.66	16	144.83	0.23	0.32	**0.28**	0.63
5 Cross level	1	N/A	0.67	2.48	**1.58**	1.58	4	2u	0.51	0.84	**0.67**	0.88

Overall, using the accommodation effect analysis ($g_{\text{relative effect}} = g_{\text{target}} - g_{\text{regular ed}}$), accommodations had a positive effect on the target population and an almost zero effect on general education students. Relatively speaking, test accommodations had a small positive effect on "target population" students (ignoring, for the moment, the differences among the various target groups), using general education students as a comparison group. The overall weighted mean accommodation effect for the all target population students was 0.16, with a standard error of 0.02 (Q = 470.97, df = 46, p < 0.01). For general education students, the overall weighted mean effect was 0.06, with a standard error of 0.02 (Q = 392.83, df = 30, p < 0.01). The relative accommodation effect (based on studies that provided empirical results for both populations), was 0.10 with a standard error 0.03 (Q = 104.14, df = 30, p < 0.01). Despite the positive accommodation effects, the significant Q test for homogeneity of variance revealed that the variations among the accommodation effects were large, implying that using the mean effect alone could be misleading because it would fail to portray the diversity of accommodation effects.

BEST COPY AVAILABLE

The results in the right side of Table 3 provide summaries of the effects by subcategories. The majority of the Q_W tests was significant ($p < 0.05$), indicating that the accommodation effects varied substantially *within* different types of accommodations, different ways of identifying target populations, and the grade levels of the students. For example, even though the average effect of *timing* was 0.31 for the target students, its associated Q_W ($\approx \chi^2_{df=21} = 126.64$) revealed that the mean effect across those 22 effect sizes was highly variable. On average, *assistive device* (0.35), *setting of test* (0.47), and *radical accommodations* (0.25) seemed to be somewhat effective in improving the scores of target population students, although each type of accommodation, except for the "combination" category (homogeneous Q_W ($\approx \chi^2_{df=4} = 3.06, p = 0.54$), exhibited great internal variability. *Presentation formats* and *response formats* exhibited small negative effects of -0.14 and -0.02, respectively.

Relative Effects — Do Accommodations Offer Target Populations a Comparative Advantage?

In order to enable parallel comparisons between the results in the right half of Table 3 (where the effects are limited to target populations) and those in the left half of that table (where the relative effects are listed), we recomputed the weighted mean accommodation effect for the target population (in the right half), using only those effect sizes from studies in which both populations were included (i.e., those within an *equivalent-groups* design or a *test-retest* design, which are those in the left side of Table 3). The mean effect size was 0.11, with a standard error 0.02 ($Q = 387.34$, $df = 30$, $p < 0.01$). Column F of Table 3 shows the relative effects (the difference between the effect on the target population and the effect on the regular population) and the comparable effects for the target population (i.e., when the target population is limited to those studies which included both target populations and regular populations) in parentheses.

Among the seven types of accommodations, the mean relative effect of *Presentation Format* was homogeneous (Q_W ($\approx \chi^2_{df=2}$ = 0.67, p = 0.72) with a weighted average of 0.09. The other types of accommodations exhibited heterogeneous effects. *Setting of test* had the largest mean relative effect at 0.68; *radical accommodations* had the second largest mean relative effect at 0.29; *assistive device* and *timing of test* had the same mean effect of 0.07; and both *combinations of accommodation* and *response format* exhibited negative effects (-0.17 and -0.11, respectively).

The small negative relative effects for *combinations of accommodation* (-0.17) and *response format* (-0.11) suggested that the general education population benefited slightly more from these accommodations than did the target populations. Because both categories of accommodation were associated with wide confidence intervals (the effects were based on only a single effect size in one empirical study in both accommodations), these results should be regarded with great caution.

We also investigated the difference of the mean relative effects and the mean effect for the target population by limiting the analysis to that subset of studies that included students from both the target and regular populations (Column F). The analysis suggested that some accommodations benefited both the target and general education populations while others benefited one population but not the other. For instance, the average effect of *timing of test* on the target population was 0.37 but relative to the effect for the general population, its effect was only 0.07 (see Cell F15). This implies that *increases in the time alloted* raised the scores of both the target and general education populations with a slightly larger effect in favor of the target populations. In a subsequent section, we will discuss each empirical study in light of finding underlying factors to help implement and improve this commonly used accommodation (*timing of test*). *Setting of test* (e.g., putting the students in a separate room or playing music for hyperactive students—see Cell F14) showed an average effect of 0.47 on the target population and a relative effect of 0.68, indicating that this accommodation raised the scores for the target population and lowered the scores for the general education population. *Radical accommodation* (Cell F12 of Table 3) and *presentation format* (Cell F11 of Table 3) also showed similar effects.

Although the average effect of *assistive device*, 0.07, seems to suggest that this accommodation did not help the target populations (the target populations did not increase their scores under accommodated condition nor did they increase their scores relative to the general education population (see cell F9 of Table 3), one should not put full faith to this interpretation. The wide confidence interval (-0.23, 0.38, inclusively) and significant homogeneity statistic (Q_W = 11.98) signals that this category may cover too broad a range of assistive devices. It includes, for example, providing glossaries to ESL students on mathematics tests, graphic organizers to LD students in social science tests, and lenses to visually impaired students. In Table 4 we try to disentangle these effects. A relative effect -0.12 was associated with providing glossaries to ESL students, and a rather large positive effect of 1.58 was associated with providing graphics organizers to LD students. This contrast underscores the specificity of the interaction between accommodations and the specific needs of different target populations. The sparse pattern in Table 4 reflects the unfortunate fact that very little work has been done on this category of accommodations; thus while the work we have is encouraging, there is so little of it that we must stress caution in interpreting results and recommending policies. On the other hand, these promising suggestions should provide us with strong motivation to pursue active research on some of these assistive devices with particular target populations (e.g., modifications of presentation format and or response format for ESL and LEP students; modifications of response format for LD students).

Table 4: Average Relative Accommodation Effect ($g_{target} - g_{regular_ed}$) by Student Status

	A	B	C	D	E	F	G	H	I	J
1	Student Status / Accommodation Effects		Assistive Devices	Combinations of Accommodations	Presentation Formats	Radical Accommodations	Response Format	Setting of Test	Time of Test	Total
2	ESL / LEP / Subgroups	Mean	-0.12	-0.17		0.29 (0.29)			-0.12	
3		SD				0.1				
4		N	N=1	N=1		N=2			N=1	5
5										
6	Students with Garden Variety Disabilities	Mean			-0.03				0.04	
7		SD								
8		N			N=1				N=1	2
9										
10	Hyperactive Students	Mean						0.68		
11		SD								
12		N						N=1		1
13										
14	LD Students	Mean	1.58		0.01	0.88 (0.26)			0.68 (0.09)	
15		SD				1.17			1.55	
16		N	N=1		N=1	N=3			N=14	19
17										
18	Students with No Formal Status	Mean			0.12		-0.11		0.05 (-0.03)	
19		SD							0.15	
20		N			N=1		N=1		N=2	4
21										
22		Total	2	1	4	4	1	1	18	31
23										
24	Note: Weighted average effect sizes are reported in parentheses.									

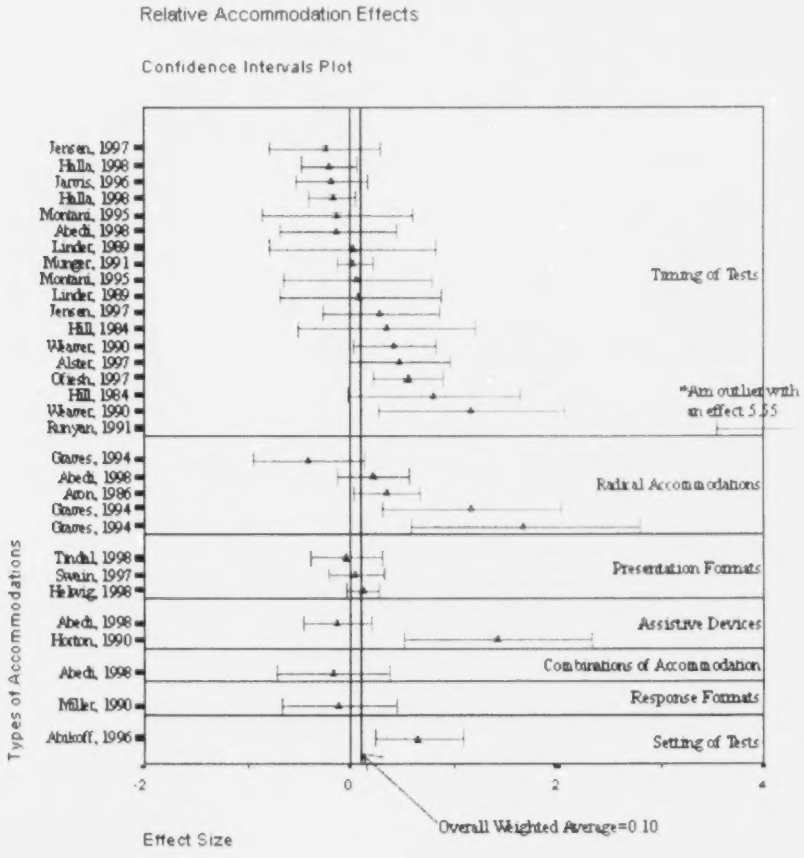
Just as it may be inappropriate to examine the interaction between accommodations and the nature of the target populations, it may also be inappropriate and difficult to investigate the interaction between the effects of accommodations and age levels. For example, some researchers have suggested that diagnosing learning disabilities is much more difficult and inaccurate for elementary students than for college students. Table 3 (Cell A27 - M31) shows the average accommodation effects by different age groups. The relative accommodation effect was slightly lower (0.09 vs 0.13) and less variable ($Q_{fl} \approx \chi^2_{df=7} = 9.81, p = 0.20$) for K-6 than for postsecondary students.

The relative accommodation effects and their confidence intervals, summarized in Figure 2 (p.20), contrast the average effects and their variability across the seven types of accommodations. As can be seen, the effects of accommodations vary even within the same type of accommodation. This variability within categories suggests one of three possible interpretations either: (1) the accommodation is unstable because of sampling errors, or (2) its effects are highly specific; that is, they emerge only in particular contexts with particular populations, or (3) we have not found the appropriate scheme for grouping and characterizing the accommodations. Because we are most interested in evaluating the possibilities of explanations 2 and 3, we further analyzed effect sizes and provided narrative descriptions

BEST COPY AVAILABLE

of the studies in order to see if other, more particular, ways of grouping the studies might shed light on why the effects emerge in some cases but not others.

Figure 2: Confidence Intervals Plots (95%) for the Relative Accommodation Effects



BEST COPY AVAILABLE

Ad Hoc Regrouping by Canonical Accommodation Effects

To help us ferret out the most plausible explanations for variable effects, we regrouped the effects into four canonical types based on the nature of the findings. These types, represented visually in Figure 3 (p. 22) specify the relationship of effects on target and regular populations. The relative accommodation effect size (if it is positive, then an advantage accrues to target population students over regular students) is plotted on the vertical axis while the effect of the accommodation for the target population only (comparing accommodated to standard testing conditions) is plotted on the horizontal axis. The diagonal line from the point (0,0) to (3,3) could be best thought of as an index of "indifference" to non-target (i.e., regular) students. At any point on that diagonal, the accommodation effect for regular students (the difference between accommodated and standard testing conditions) is zero: that is, they are neither helped nor hindered by the accommodation.

Interpreting relative accommodation effects ($g_{\text{relative effect}} = g_{\text{target}} - g_{\text{regular ed}}$) requires extra caution because a positive effect merely indicates that the average score change for target students (g_{target}) is larger than that for regular students ($g_{\text{regular ed}}$). Positive relative accommodation effects can result from two very different scenarios. In the first scenario, depicted in Region I of Figure 3, which we might label as *beneficial for the target population and disadvantageous for regular students*, the regular population actually does worse in the accommodated versus the standard condition. In Region II, we might label *beneficial for all students but more beneficial to target students*, the accommodation increases the scores of target students but does not decrease the scores of regular students or it increases their scores to a lesser extent than the increase that accrues to target students.

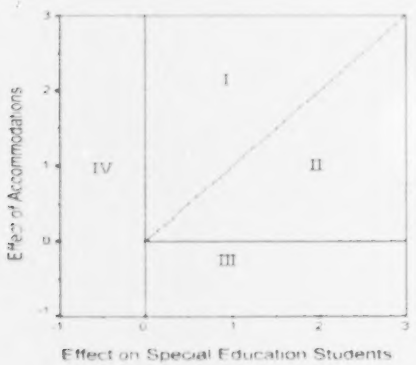
Interpreting a negative relative accommodation effect requires as much caution as interpreting a positive relative accommodation effect. The scenario depicted in Region III represents those effects that we might label *beneficial for all students but less beneficial to target students* (i.e., both target and regular students improved their scores in the accommodated condition but regular students benefited to a greater

extent). In another scenario, depicted in Region IV, we can identify effects that are *disadvantageous to target students*. Effects depicted in this region are associated with accommodations that lowered the scores for target students.

The interpretations of the relative effect sizes become explicit with the following examples. A relative accommodation effect would be plotted above the diagonal line if an accommodation has lowered the scores for regular students while improving the scores for target students. A hypothetical point (1,2), located in Region I, illustrates this scenario — a beneficial accommodation effect for target students ($g_{target} = 1$) but a disadvantageous effect for regular students ($g_{regular\ ed} = g_{target} - g_{relative\ effect}$; 1 - 2 = -1). A relative accommodation effect would be plotted below the diagonal line if an accommodation has improved the scores for both regular and target students but to a greater extent for target students. This scenario is represented by another hypothetical point (3,2), located in Region II, which represents a beneficial accommodation effect for all students ($g_{target} = 3$; $g_{regular\ ed} = 3 - 2 = 1$) but the accommodation is more beneficial for target students.

A third hypothetical point (-1, 2), located in Region IV, shows that the accommodation effect was disadvantageous to both target students ($g_{target} = -1$) and regular students ($g_{regular\ ed} = -1 - 2 = -3$). Notice that all three aforementioned hypothetical points represent the same relative accommodation effect ($g_{relative\ effect} = 2$) even though they represent a very different effect for the regular and target students.

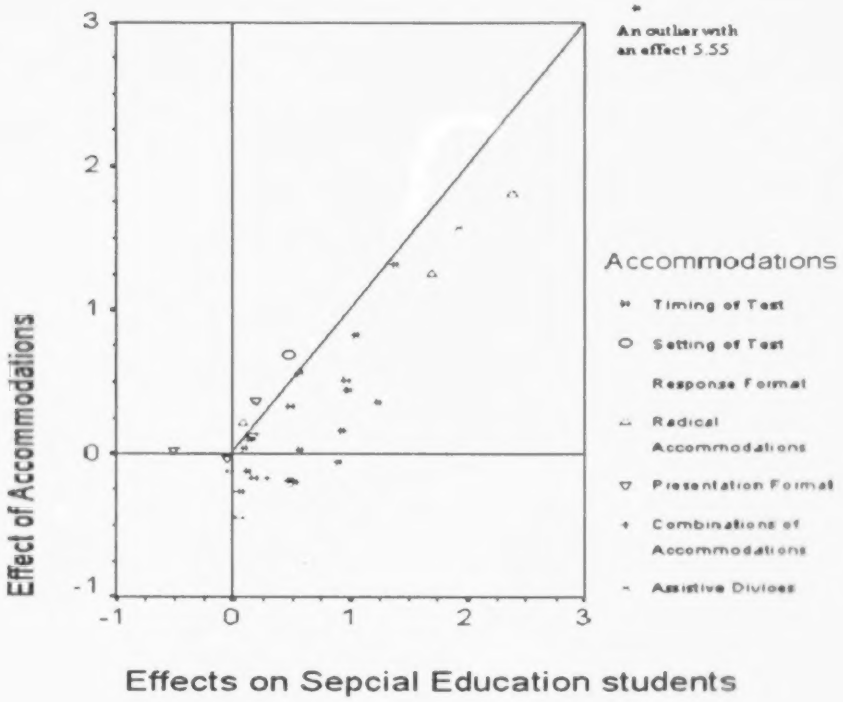
Figure 3: Prototype of Four Canonical Accommodation Effects



BEST COPY AVAILABLE

In the section that follows, we examine the specific characteristics of specific studies that may very well determine the conditions under which a particular accommodation, such as timing, may or may not prove to exhibit optimal effects.

Figure 4: Canonical Accommodation Effects and Types of Accommodations



BEST COPY AVAILABLE

Table 5: Two Types of Effects for the Extended-Time Accommodation

	A	B	C	D	G	H	I
1		ID		Research Studies	Relative Accommodation Effects	Special Education Status	Grade Level
2			1	Alster (1997)	0.492	4) LD	4) Postsecondary
3			2	Munger (1991)	0.04	2) Garden Variety	1) K-6
4			3	Runyan (1991)	5.548	4) LD	4) Postsecondary
5			4	Hill (1984)	0.812	4) LD	4) Postsecondary
6			5	Hill (1984)	0.353	4) LD	4) Postsecondary
7		Region II: More Beneficial to Target Students	6	Montani (1995)	0.073	7) No Formal Status	1) K-6
8			7	Jensen (1997)	0.296	4) LD	4) Postsecondary
9			8	Ofiesh (1997)	0.555	4) LD	4) Postsecondary
10	Timing of Test		9	Weaver (1993)	0.427	4) LD	4) Postsecondary
11			10	Weaver (1993)	1.172	4) LD	4) Postsecondary
12			11	Linder (1989)	0.032	4) LD	4) Postsecondary
13			12	Linder (1989)	0.095	4) LD	4) Postsecondary
14				*Subtotal (N)*	12	12	12
15			1	Halla (1988)	-0.165	4) LD	4) Postsecondary
16			2	Halla (1988)	-0.195	4) LD	4) Postsecondary
17		Region III: Less Beneficial to Target Students	3	Montani (1995)	-0.132	7) No Formal Status	1) K-6
18			4	Jensen (1997)	-0.242	4) LD	4) Postsecondary
19			5	Jarvis (1996)	-0.183	4) LD	4) Postsecondary
20			6	Abedi et al (1998)	-0.122	1) ESL / LEP / Cultural Subgroup	2) Middle School
21				*Subtotal (N)*	6	6	6
22				*Total N*	18	18	18

What Makes *Timing of Test* Less Beneficial to the Target Population than to the General Education Population?

Using the computerized version of the Nelson-Denny Reading Test (NDRT), Jensen (1997) reported that generally overall college students without LD tended to benefit more than students with LD under an unlimited time accommodation. In addition, Jensen examined two factors contributing to this finding, namely the order effect and the test-type effect. The order effect was examined by manipulating the order of the timing condition (whether the unlimited time condition occurred first or second) provided to groups of students with and without LD. The untimed condition was more beneficial to LD students than to non-learning disabled (NLD) students when it was administered first. However, when it was administered second, LD students did not benefit from the additional time (g_{target} = -0.69, Table 5 on page 24). This order effect suggested that LD students might have been fatigued due to the extended testing (Jensen, 1997; p.63). The performance of NLD students did not tend to be influenced by the order effect. In Jensen's account, such an order effect on LD students implies that additional time could be beneficial

BEST COPY AVAILABLE

to LD students only if the additional time was not so long as to distract them. Other than the order effect, Jensen reported that the benefit unlimited time provided to LD students depended on the cognitive complexity of the test. Specifically, the effect of providing unlimited time to LD students is more profound for literal-type questions than for implicit-type questions. Jensen noted that unlimited time appeared to be more effective to LD than to NLD students when it was provided for tests not requiring complex cognitive abilities. Jensen also asserted that the effects of unlimited testing time to LD students was sensitive to specific factors like those just mentioned and there is no justification for the generalization of the effects to other situations.

In another study, Halla (1988) found that relative to NLD college students, LD college student might not necessarily use and benefit from unlimited testing time. Halla reported that NLD students benefited more from the unlimited time than did the LD students. Students in both ability groups were tested on the GRE test and the Nelson Denny Reading Test (NDRT). In both tests, the effects of unlimited time was lower for LD than for NLD students ($g_{relative\ effect}$ = -0.17 and -0.20 for the GRE and NDRT, respectively. See Table 5). Despite the unlimited time provided, on average, students with learning disabilities used less time than the NLD students in the GRE test (197 vs 218 minutes) and only slightly more time than the NLD students in the NDRT (49.6 vs 48.3 minutes) (p.143). Consistent with Jensen's (1997) conclusion, Halla suggested that LD students might not necessarily be able to take advantage of the extended time to demonstrate their optimal ability. Specifically, LD students did not seem to benefit from extra time offered to them when the standard testing time was long (e.g., three hours for the GRE test).

In a third grade arithmetic test, Montani (1995) reported that providing unlimited time increased the scores of all students — students with no disability status (NLD) and students with low reading ability but normal mathematics ability (presumably LD students). Montani suggested that the LD students benefited on both number-fact problems and story problems (g_{target} = 0.93 and 0.89). Although these effects appeared to be large compared to other studies discussed in the current report, these effects were not as large when compared to the effect on NLD students. For number-fact problems, the effect of

unlimited testing time on LD students was only slightly greater than that of the NLD students ($g_{\text{relative effect}}$ = 0.07. See Table 5). For story problems, the unlimited time effect was less profound for LD than for NLD students ($g_{\text{relative effect}}$ = -0.13. See Table 5). These relative results appeared to indicate that students with and without learning disabilities both benefited from unlimited time. Put differently, such results suggest that the test was a *speeded* test; that is, a test for which the standard testing time might not be sufficient to NLD students, in addition to LD students. Students were allowed only three seconds to respond to a question regarding basic number-fact as well as to story problems. Responses were considered correct only if they provided a correct answer within the time limit following the experimenter's reading of the question. Based on the this evidence, elementary students (third graders) without disabilities need more than three seconds to respond to questions as straightforward as simple number-fact questions. Students with learning disabilities might need even more than that.

In an on-going study, Abedi, Hofstetter, Baker, and Lord (1998) used the eighth-grade mathematics examination in the National Assessment of Education Progress (NAEP) to investigate the extended-time effects on students with Limited English Proficiency (LEP) and students with Fluent English Proficiency (FEP). Seventy-two percent of the target population was Hispanic. Abedi et al (1998) found that all students benefited from receiving extended-time, but the LEP students did not benefit as much as the FEP students ($g_{\text{relative effect}}$ = -0.12. See Table 5). Abedi et al. determined that reading level was not a confounding factor because the differential benefits still existed even after they controlled for students' reading levels. Although these results indicated that the test was not a pure power test, the finding that FEP students increased their scores by only 0.26 standard deviation suggested that the speed factor was minimal.

Using a classroom test (business and hospitality) administrated to college students, Jarvis (1996) found that LD students were more aware of and precise regarding their needs for additional time than were NLD students. They also found that all students benefited from receiving unlimited time. The timing effect for LD students was lower than that for the NLD students (g_{target} = 0.50 & $g_{\text{regular ed}}$ = 0.69). The unlimited-time was given to students on a need basis and the time used by the students ranged from 1 to

70 minutes. In a posttest questionnaire, all LD students reported that they typically used extra time on tests. Even though the majority (91%) of NLD students reported that they did not need extended time and it would not help, NLD students as a group benefited more than did LD students.

What Makes an Extended-Time and an Unlimited-Time Accommodation Work?

In a recent study, Alster (1997) investigated the effect of providing unlimited time to postsecondary students on an algebra test (ASSET is a test developed by ACT). She recorded the time it took the students to complete the test under the unlimited time condition and concluded that the LD students used more time (approximately 13 extra minutes on a test designed to be completed in 12 minutes), on average, than did NLD students (who used approximately 8 more minutes). In addition, the time used by individual LD students was more variable than that used by NLD students (the range was 12 to 56 minutes for LD and 12 to 31 for NLD). Unlike Halla (1988) and Jensen (1997), Alster found a substantial accommodation effect for the LD students as well as a relative accommodation effect favoring LD students. There was also an order effect for both LD and NLD students, with better performance noted on the untimed condition when it followed the timed condition, although this *practice* effect was more pronounced for the LD than for the NLD students. Averaging across the two ordered conditions, providing unlimited time to students appeared to be more beneficial to LD than to NLD students ($g_{\text{relative effect}} = 0.492$. See Table 5). In commenting on the *practice* effect, Alster argued that LD students might need more practice in order to overcome their anxiety or previous failures when taking tests under a time limit.

Even though Halla (1988) and Jensen (1997) found that administering the extended-time condition favored LD students, their findings can be viewed as supplementing rather than opposing Alster's findings. Halla and Jensen suggested that extending time when the original time limit may already be quite long (e.g., one to three hours) may not benefit either LD or NLD students, citing an

overall fatigue effect as washing out any potential benefit of extended time. They did not, however, study tests of shorter duration. By contrast, Alster found that the extended time benefited both groups, with an extra benefit accruing to the LD students, when the original testing time allocation was short, i.e., 12 minutes. Parenthetically, it should be added that Alster suggested adding practice items at the beginning of an exam to lower the anxiety level of all students, but especially LD students.

Ofiesh (1997) administered the Nelson Denny Reading Test to students in postsecondary school to examine the *timing of test* effect. Instead of an unlimited time condition, Ofiesh used a condition with a fixed additional amount of time (60% more time to every examinee) and found differential timing effects for LD and NLD students. Providing a fixed amount of extra time seemed to be practical and effective ($g_{\text{relative effect}} = 0.56$. See Table 5) in accommodating students; in fact, the pattern found by Ofiesh matches the ideal scenario described in Figure 3; the target population students benefited from the accommodation ($g_{target} = 0.57$) while the regular education students were neither advantaged nor disadvantaged ($g_{\text{regular ed}} = 0.01$). Regardless of the fact that timing conditions were not counterbalanced, the practice effect seemed to be minimal, at least to the NLD students, because the mean scores of the NLD students did not increase even after they practiced in the timed condition. The findings in the work of Ofiesh (1997) and Alster (1997) suggest a practical and effective solution to accommodate the needs of special education students, that is, to provide a fixed amount of additional time (e.g., by a factor of 1.5 to 2.0 of the standard time) and a number of trial questions that are not counted as part of the official score.

What are the effects of providing additional time on a need basis? Runyan (1991) used the Nelson Denny Reading Test to examine the effect of providing additional time to college students who could not complete the test under the timed condition. All LD students (n=16) did not complete the test on time, whereas the majority of the NLD students (87% out of 15 students) completed the test within the 20-minute standard time allotment. The LD students increased their scores from the low tenth percentile to the mid-seventieth percentile when additional time was provided. Such an enormous advantage for LD students over NLD students in the accommodated condition ($g_{\text{relative effect}} = 5.55$. See Table 5) deserves explanation, and the likely explanation points to an artifact of the test administration conditions. Unlike

the other aforementioned studies, which used two parallel test forms to examine the timing effect, Runyan (1991) used a "red line" procedure. In this procedure, students are asked to draw a red line under the last item completed during the allocated time, but then to continue to respond to items, during the extended time, not previously encountered in the standard time. Runyan, however, did not allow students to return to previously encountered items to change answers (i.e., those completed in the timed phase). While this procedure may have avoided measurement errors (due to sampling test items) by employing a single test form, it might have introduced some artificial experimental noise by not allowing students to return to previously encountered items to change their answers. These results might be construed as inconsistent with the findings of Halla, Jensen, and Alster (reviewed earlier) because of the fact that additional time proved beneficial to the LD students even when the originally allocated time limit was relatively long (i.e., three hours), bringing the fatigue explanation into question. An alternative explanation of these particular results is that the test was so difficult for this particular sample of students (recall that none of the LD students finished the test in the standard time allocation) that extra time outweighed the disadvantageous effect of fatigue.

Using a similar procedure (drawing a line under the last item finished in the standard time allocation), Weaver (1993) determined that reading level and institutional factors influenced LD students' advantage using the timing effect on the Nelson Denny Reading Test. She hypothesized that LD students with relatively lower reading scores might not be able to take as much advantage of the extra time as those with higher reading level. LD students with low reading levels were recruited from Community Colleges while LD students with high reading levels came from four-year universities in her region. Learning disabled students in both settings were compared to an NLD group in their corresponding institution. Weaver found that LD students in the community college setting benefited less ($g_{\text{relative effect}}$ = 0.43. See Table 5) than LD students in the university setting ($g_{\text{relative effect}}$ =1.17).

In addition to comparing the reading level, community college-versus-university effects, and extended-time effect for the students, Weaver also examined the additional value of "optmizing" the test-taking setting. She compared students from both institutional settings who had received extended time

(those included in the earlier comparsion) with another set of students who received unlimited time *and* took the test in a quiet setting (test administered on an individual basis in a quiet room). She concluded that this optimal setting (unlimited time in a quiet setting) yielded a greater relative effect than extended-time alone, in which the extended-time test was administrated in a group setting.

Munger and Loyd (1991) wanted to determine whether target students (94% LD and 6% physically handicapped) could and should be included in standard testing conditions. The fifth grade participants responded to the Language Usages and Expression and Mathematic Concepts subtests in the Iowa Test of Basic Skills (ITBS). Comparing an unlimited time condition to a timed condition, Munger and Loyd found that both the LD and NLD groups showed only a slight increase in their scores even though both groups were given as much time as they wanted to use. Relative to the NLD students, LD students benefited only very little from the unlimited time condition ($g_{\text{relative effect}} = 0.04$. See Table 5). Munger and Loyd noted that the ITBS was designed to be a power test in which sufficient time has been allowed to all students to attempt all the test questions. If the standard time condition already, in a sense, accommodated the LD students, additional time should not help. In fact, when comparing to the Nelson Denny Reading Test employed by the other studies discussed, the ITBS seemed to provide more time per question. Students have 30 minutes to answer 38 questions in the language usage and expression test (about 47 seconds per item) and 25 minutes to answer 35 items (42 seconds per item) in the mathematics test (p.55). The Nelson Denny Reading Test allows 19 minutes for 38 questions (30 seconds per question) in the comprehension section and 15 minutes for 100 items, each with five answer choices (9 seconds per question), in the vocabulary section (Hill, 1984; Linder, 1989; & Weaver, 1993).

Having examined several studies that employed the Nelson Denny Reading Test, we wondered whether it is really a speeded test, especially its vocabulary section. Of the six studies that used the NDRT, three studies used the reading comprehension and vocabulary sections and they all reported that even NLD students benefited from receiving additional time. Hill (1984) reported an effect of 0.89; Linder (1989) reported an effect of 0.54; and Weaver (1993) reported effects of 0.55 and 0.21 for university and community college students, respectively. These three studies administered all 100 items in

the vocabulary section. Conversely, Ofiesh (1997) reduced the length of the vocabulary section by 20 items (25% of the length used in the other studies) and reported that NLD students almost did not benefit from the timing effect (0.01). In addition, Runyan (1991) did not administer the vocabulary section and reported that NLD students benefited only very little (0.1) from the timing effect.

Hill (1984) administered both the ACT and the Nelson Denny Reading Test to investigate the timing effect. Hill found that the average additional time used by the LD students was only slightly higher than that used by the NLD students (23.7 minutes versus 20.2 minutes), indicating that the standard time might be insufficient even for NLD students. Given unlimited time, LD and NLD students benefited substantially, exhibiting gains of 1.24 and 0.89 standard deviations, respectively. When the ACT was used, the LD students still benefited by a great extent from the unlimited time condition (1.04 S.D.), yet the NLD students only benefited by 0.23 standard deviation. Based on our assumption that the NDRT is a speeded test, we considered the relative accommodation effect (0.35 = 1.24 - 0.89) to be conservative. With the evidence based on the ACT assessment, which provided sufficient time for the NLD students in the standard condition, we are inclined to believe that the relative accommodation effect could be as high as 0.81 (see Table 5).

Among the studies discussed so far, only two studies examined the timing effects on elementary students (K-6); both found a small relative timing effect (Montani, 1991, $g_{\text{relative effect}} = 0.073$; Munger and Loyd, 1991, $g_{\text{relative effect}} = 0.04$. See Table 5). Although those small relative timing effects might lead one to think that extra time or unlimited time should not be provided to young children with learning disabilities, such a conclusion is not warranted. Whether or not additional time can accommodate students' needs depends on the length of the test and the standard time allowed. In the Montani (1991) study, the small relative effect may have been due to the fact that the standard time was insufficient for *both* NLD and LD students. The fact that all students benefited from extended time implies that more time should be added to the standard condition for all students.

Results from Large Scale Studies

In addition to studies that are set out to examine test accommodation effects, we found a few studies that, while they could not be included in our meta-analysis (usually because they did not report sufficient statistical data or because they were unable to establish appropriate control groups), do add to our understanding of the impact and appropriateness of various accommodation practices. Most often, these are studies that examine the effects of accommodations in large-scale examinations. In a review, Bennett & Ragosta (1985) reported that extended-time was beneficial to examinees with disabilities. Students with physical disabilities benefited more than LD students in both the verbal and math sections, relative to the comparison group of students without disabilities. LD students improved their scores as much as did the physically disabled students, although the LD students' accommodated scores were still lower than their NLD comparison group in the standard condition (-0.27 & -0.12 SDs for verbal and math, respectively).

In another study using the SAT test, Bennett & Rock (1989) found consistent results indicating that students with disabilities (visually impaired) accommodated scores (i.e., when given extra time) were lower than non-disabled comparison group (when tested in standard time conditions). Given extended-time, visually impaired students scored almost as high as students without disabilities on the verbal section but still scored considerably lower than students without disabilities on the math section.

Statewide studies were another source for evaluating accommodation effects. Shepard, Taylor, Betebenner, and Weston (in press) examined the predictive validity and effects of various accommodations provided to special education and LEP students. They compared student relative standing (using z scores) on the Metropolitan Achievement Test (MAT) test with their relative standing on Rhode Island's Grade 4 Mathematics Performance Assessment. The performance assessment provided accommodations and aimed at including as many students as possible, whereas the MAT did not provide accommodations. Using the MAT scores for comparison, Shepard et al. found that both LEP and special

education students improved their z-scores when they were accommodated in the performance assessment. Of the accommodated students, the relative *z*-score improvement was 0.51 for LEP students (with greater than two years of English speaking experience) and 0.50 for students in special education placement (for 50% or more of the class time). Of the students who were not accommodated, the *z*-score improvement was small for both LEP and special education students. Shepard et al. (in press) asserted that the relative advantage on the accommodated performance assessment should be interpreted with caution because accommodations might not be provided in a standardized fashion across the many participating schools. Even so, their results point in the same general direction of improvement for target populations when the extra time is really needed, as one might expect it to be in a performance examination.

CONCLUSIONS, RECOMMENDATIONS, AND FUTURE RESEARCH OPPORTUNITIES

Among the target populations of special education and limited English proficient students throughout the nation, students with learning disabilities was the subgroup for which we found the largest number of empirically based studies examining the effects of test accommodations. Providing *extended-time* or *unlimited time* is by far the test accommodation of choice — the accommodation for which abundant empirical research could be found. Research efforts on other accommodations such as *assistive devices*, *combinations of accommodation*, *presentation formats*, *response formats*, *setting of tests*, and *radical accommodations* were so rare that we hesitate to even pose tentative conclusions about their effects. Additional research studies that allow us to understand and predict the effects of various *presentation formats* and *response formats* in relation to particular target populations deserve much attention.

We found that a tremendous amount of variation existed in each category of the aforementioned accommodations, indicating a high degree of effect specificity; the variation of effects between each category was as large as the variation within each category. Specifically, sometimes target students benefited from accommodations, sometimes they did not, and sometimes they were actually disadvantaged by the accommodation.

Providing an accommodation to "all students" alone did not emerge as the best resolution for the controversial issue in determining who should be accommodated; our findings indicated that improperly implemented accommodations could in fact be disadvantageous to either the target population or the regular education population. Specifically, our findings showed that providing accommodation(s) to regular education students might lower their test scores when they did not need to be accommodated; this was especially true for accommodations other than extended time (e.g., *setting of test*, *presentation format*, and *radical accommodations*—see region I in Figure 4).

Based on our investigations on the *extended-time* or *unlimited-time* accommodation, we found, overall, that extended time is beneficial to target populations when compared to standard time conditions (average effect size (g_{target} = .37). Regular education populations also benefit from extra time ($g_{regular\ ed}$ = .30); however, the comparative advantage to target populations is quite modest ($g_{relative\ effect}$ = .07).

What is more important about the corpus of studies on extended time is the tremendous variation in observed effects among studies. Even so, there are two factors that might explain the wide variations of this particular accommodation: 1) the constitution (and appropriateness) of "standard conditions" and 2) the nature of specific implementations of the accommodation. Based on our meta-analysis and narrative analysis, we found that some tests, such as the Nelson Denny Reading Test, may well have such tight standard conditions (30 seconds or less per question in a comprehension reading test and nine seconds per question in a vocabulary test) that even regular education students seemed to benefit from receiving additional time. And in those instances, the relative effects tended to be quite small or even negative. By contrast, in tests, such as the ASSET or a modified version of the Nelson Denny, with more generous standard time allocations, target populations tended to benefit from increased time. In our

review of the literature, we found evidence indicating that target populations might benefit from receiving a brief period with "practice items" to lower their anxiety level and to acquaint themselves with the test. At the other end of the continuum, some tests are so long that all extended time does for target students is to allow fatigue and distractions to set in.

We think these two factors (standard conditions and the implementation of the accommodation) might form the basis of some guidelines for determining a proper way to administer accommodations.

Suggested Guidelines

A) Make sure that standard conditions are appropriate for regular education populations. This principle could be implemented via the following steps:

Step 1: Determine the baseline or reference condition (e.g., time limit) in which most or all regular students could complete a given test.

Step 2: Provide accommodations (e.g. extended time or unlimited time) for the regular students and compare their scores in the accommodated condition with their scores in the baseline condition.

Step 3: If the scores of regular students have increased notably from the baseline condition to the accommodated condition, then reconsider altering the baseline condition (e.g., increasing the standard time limit).

Step 4: Repeat this comparison procedure between the baseline and accommodation conditions until the increment of test scores becomes no longer noticeable.

B) Provide feasible and valid accommodations (which should not alter the underlying construct being measured) tailored to the needs of the target populations. To determine whether or not specific accommodations alter the construct(s) measured in a test, one may conduct task analyses to understand the tasks involved. For example, cognitive analysis (e.g., think aloud analysis) could be used to discover the characteristics of *presentation format* or *response format* that could help the target populations perform at their optimal level while retaining the construct validity of tests (Pearson, 1998).

Another issue, not examined in the current study due to time and space limitations, includes investigation of the criterion-related validity (Crocker and Algina, 1986) of accommodations provided to the target population and the regular education population. Understanding the relative standing of students in both standard and accommodated conditions is invaluable for improving the educational opportunities for all students, target students in particular (e.g., would accommodations change the predictions of college successfulness?). Research geared toward this trend is exemplified in the work of Shepard et al. (in press).

In the current research study, we used the *average relative effect size* to determine the effects of accommodations. This index capitalized on the mean effect of accommodation and helped address the question of whether target students, as a group, improve their test scores from the standard condition. If one is more interested in knowing whether or not an accommodation has *diversified* or *homogenized* the scores of the target populations, one might examine the *dispersion* effect (Chiu, 1999; Harwell, 1997; Kalaian and Becker, 1996; Raudenbush, 1988) of accommodation. Examining the dispersion effect allows one to address such questions as, Does an accommodation change the score variability of the students?

Given the proliferation of empirical research in recent year (Olson and Goldstein, 1997), particularly the increasing amount of research conducted in many states, a cross-state or multiple site meta-analysis may be useful to account for current inclusion efforts and results, which in turn could guide the decisions surrounding the design and implementation of accommodations.

Swap labels

APPENDIX A: Summary of Studies Analyzed*.

ID	Authors	Year	Age	Accommodation Used	Target Students	Research Design	Type of Outcome	Test Used	Effect Regular Ed	Effect Target Students	Relative Effect	N
1 A	Abedi, Hofstetter, Baker, & Lord	1998	Middle	Radical Accom	ESL / ELP / Cultural Subgroups	Equivalent Group on Target and Regular Students	Math	Math	0.08	-0.14	0.22	275
1 B				Combinations				Math	0.29	0.46	-0.17	179
1 C				Time				Math -- Extra Time	0.13	0.26	-0.12	180
1 D				Devices				Math -- Glossary	-0.05	0.07	-0.12	299
2 A	Abikoff, Courtney, & Harold	1996	K - 6	Setting	Hyperactive	Repeated Measure on Target & Regular Students	Math	Arithemtic	0.49	-0.20	0.68	40
3 A	Alster	1997	Postsec	Time	LD	Repeated Measure on Target & Regular Students	Math	Math	1.00	0.49	0.51	44
4 A	Aron	1986	Postsec	Radical Accom	ESL / ELP / Cultural Subgroups	Repeated Measure on Target & Regular Students	Reading	New Jersey college basic skills placement test	0.19	-0.17	0.37	62
5 A	Bennett, Rock, & Kirele	1987	Postsec	Combinations	Mixed Groups	Equivalent Group on Target & Regular Students	Math	Quantitative	0.04	na	na	339
5 B							Verbal	GRE Verbal	0.23	na	na	339
6 A	Blaskey, Scheiman, Parisi, Ciner, Gallaway, & Selznick	1990	Cross Level	Devices	Visually Impaired	Repeated Measure on Target Students	Reading	Stanford Reading Test	0.09	na	na	22
7 A	Braun, Ragosta, & Kaplan	1986	Postsec	Combinations	Mixed Groups	Equivalent Group on Target & Regular Students	Math	GRE-Quantitative	0.14	na	na	203
7 B							Verbal	GRE-Verbal	0.08	na	na	289
7 C							Verbal	GRE-Verbal	-0.11	na	na	203
7 D							Math	GRE-Quantitative	0.02	na	na	289

Note: Numeric IDs indicate a research study and alphabetical IDs indicate multiple effect sizes within the same study.

BEST COPY AVAILABLE

Accommodations

ID	Authors	Year	Age	Accommodation Used	Target Students	Research
8 A	Cauderv	1990	High Schl	Time	ESL / LEP / Cultural Subgroups	Repeated Target St
9 A	Chen & Graves	1995	Postsec	Radical Accom	ESL / LEP / Cultural Subgroups	Equivalen Target & Students
9 B						
10 A	Gallina	1989	Cross Level	Time	Physically Handicapped	Repeated Target St
11 A	Graves, Semmel, & Gerber	1994	Middle	Radical Accom	LD	Repeated Target & Students
11 B						
11 C						
12 A	Halla	1988	Postsec	Time	LD	Repeated Target & Students
12 B	Halla	1988	Postsec	Time	LD	Repeated Target & Students
13 A	Helwig, Rozek-Tedesco, Heath, Tindal, & Almond	1998	K - 6	Presentation	No Formal Status	Repeated Target & Students
14 A	Hill	1984	Postsec	Time	LD	Equivalen Target & Students
14 B						

Design	Type of Outcome	Test Used	Effect Regular Ed	Effect Target Students	Relative Effect	N
Measure on udents	Writing	Customized made writing test (GCE O-level)	0.15	na	na	48
t Group on Regular	Reading	Multiple-choice Test	0.09	na	na	122
		Short-Answer Test	0.49	na	na	122
Measure on udents	Math	Metropolitan Achievement Test 6 Mathematic Subtes	0.66	na	na	54
Measure on Regular	Writing	Writing – Middle prompt	0.05	0.48	-0.43	24
		Writing – Begin prompt	2.56	0.75	1.81	24
		Writing – End prompt	1.82	0.56	1.26	24
Measure on Regular	Math & Reading	GRE General test	0.18	0.35	-0.17	132
Measure on Regular	Reading	Nelson-Denny Reading Test	0.53	0.73	-0.20	132
Measure on Regular	Math	Math	0.16	0.04	0.12	247
t Group Regular	Math & Reading	ACT Total Raw Score	1.06	0.23	0.83	48
	Reading	Nelson-Denny total raw scores	1.26	0.90	0.36	48

Accommodations

ID	Authors	Year	Age	Accommodation Used	Target Students	Research D
15 A	Horton, Lovitt, & Bergerud	1990	Cross Level	Devices	LD	Repeated M Target & R Students
16 A	Jarvis	1996	Postsec	Time	LD	Repeated M Target & R Students
17 A	Jensen	1997	Postsec	Time	LD	Repeated M Target & R Students
17 B						
18 A	Linder	1989	Postsec	Time	LD	Equivalen Target & Students
18 B						
19 A	MacArthur & Haynes	1995	High Schl	Radical Accom	LD	Repeated Target S
20 A	Miller	1990	K - 6	Response	No Formal Status	Equival Target & Student
21 A	Montani	1995	K - 6	Time	No Formal Status	Repeat Target Student
21 B						
23 25	Munger & Loyd	1991	K - 6	Time	Garden Variety	Repeat Target Studer
22 B						

BEST COPY AVAILABLE

Design	Type of Outcome	Test Used	Effect Regular Ed	Effect Target Students	Relative Effect	N
Measure on Regular	Others	Social studies and science	2.08	0.50	1.58	22
Measure on Regular	Others	Classroom tests	0.50	0.68	-0.19	74
Measure on Regular	Reading	NDRT – Implicit	0.07	0.33	-0.26	20
		NDRT – Literal	0.52	0.20	0.33	20
nt Group on Regular	Math & Reading	SAT Total Raw Score	0.14	0.04	0.10	50
	Reading	NDRT Total Raw Score	0.58	0.54	0.03	50
d Measure on tudents	Science	Biology Test	0.50	na	na	20
ent Group on & Regular s	Reading	Peabody Vocabulary Test	-0.02	0.09	-0.11	96
ed Measure & Regular ts	Math	Complex math problems	1.00	1.06	-0.06	16
		Simple math problems	1.04	0.88	0.16	16
ed Measure on & Regular nts	Language	Iowa Test of Basic Skills (Language Usage)	0.12	0.06	0.07	104
	Math	Iowa Test of Basic Skills (Math Concepts)	0.08	0.07	0.01	104

Accommodations

ID	Authors	Year	Age	Accommodation Used	Target Students	Research
23 A	Ofiesh	1997	Postsec	Time	LD	Repeated Target & Students
24 A	Robinson & Conway	1990	Cross Level	Devices	Visually Impaired	Repeated Target St
25 A	Ross	1995	Postsec	Time	ESL / LEP / Cultural Subgroups	Repeated Target St
25 B						
26 A	Runyan	1991	Postsec	Time	LD	Repeated Target & Students
27 A	Swain	1997	K - 6	Presentation	LD	Repeated Target & Students
27 B						
27 C						
27 D						
28 A	Tachibana	1986	Postsec	Combinations	LD	Repeated Target St
28 B						
28 C						

Design	Type of Outcome	Test Used	Effect Regular Ed	Effect Target Students	Relative Effect	N
Measure on Regular	Reading	Nelson Denny Reading Test	0.58	0.01	0.57	60
Measure on udents	Reading	Neale Analysis of Reading Ability – Comprehension	0.90	na	na	88
Measure on udents	Writing	Pre-Professional Skills Test for Teacher Education (Group 1)	0.36	na	na	184
		Pre-Professional Skills Test for Teacher Education (Group 2)	0.13	na	na	184
Measure on Regular	Reading	Nelson-Denny Reading Test	5.95	0.10	5.84	32
Measure on Regular	Math	Keymath-R/ CAMT (Multiplication)	-0.38	-0.50	0.12	114
		Keymath-R/ CAMT (Addition)	-0.48	-0.76	0.28	114
		Keymath-R/ CAMT (Substraction)	-0.43	-0.36	-0.07	114
		Keymath-R/ CAMT (Division)	-1.20	-0.89	-0.31	114
Measure on udents	Reading	Reading Comprehension Test on DTLS (Group 1)	0.29	na	na	34
			-0.52	na	na	34
		Reading Comprehension Test on DTLS (Group 3)	0.94	na	na	22

ID	Authors	Year	Age	Accommodation Used	Target Students	Researc
29 A	Tindal, Heath, Hollenbeck, Almond, & Harniss	1998	K - 6	Presentation	Garden Variety	Repeate Target & Student
29 B						
30 A	Weaver	1993	Postsec	Time	LD	Repeate Target & Student
30 B						
					Large-Scale Studies	
31 A	Bennett & Ragosta	1985	High Schl	Time		n a
32 A	Bennett & Ragosta	1989	High Schl	Time		n a
33 A	Shepard, Taylor, Betebenner, & Weston	1998				n a

Note: Each row in the above table represents an effect size. Effect sizes with the sam
research study. The Busk and Marascuilo (1987) procedure was used to determine w
extracted from the same study. Study characteristics omitted in the table were identi

h Design	Type of Outcome	Test Used	Effect Regular Ed	Effect Target Students	Relative Effect	N
d Measure on & Regular s	Math	Math	-0.06	0.00	-0.06	213
	Reading	Reading	-0.05	-0.05	0.00	213
d Measure on & Regular s	Reading	Nelson-denny Reading Test -- Total Scores	1.56	0.24	1.32	16
	Reading	Nelson-denny Reading Test -- Total Scores	1.00	0.56	0.44	62
s						
	Verbal & Math		na	na	na	
	Verbal & Math		na	na	na	
			na	na	na	

ne numerical but different alphabetical index were extracted from the same
Whether or not multiple effect sizes should be pooled when they were
cal to those reported immediately above them.

REFERENCES[1]

*Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (1998). Validity and comparability in NAEP math assessment accommodations for students with Limited English Proficiency : UCLA/ CRESST.

*Abikoff, H., Courtney, P. J., & Harold, S. K. (1996). The effects of auditory stimulation on the arithemtic performance of children with ADHD and nondisabled children. Journal of Learning Disabilities, 29(3), 238-246.

*Alster, E. H. (1997). The effects of extended time on algebra test scores for college students with and without learning disabilities. Journal of Learning Disabilities, 30(2), 222-227.

*Aron, H. (1986). The influence of background knowledge on memory for reading passages by native and nonnative readers. TESOL Quarterly, 136-140.

Bangert-Drowns, R. L. (1993). The word processor as an instructional tool: A meta-analysis of word processing in writing instruction. Review of Educational Research, 63(1), 69-93.

Becker, B. J. (1988). Synthesizing standardized mean-change measures. British Journal of Mathematical and Statistical Psychology 41: 257-278.

Bechard, S. (1997, June 17, 1997). Overview of the state strategies for special education accommodations. Paper presented at the National Large-Scale Assessment Conference organized by the Council of Chief State School Officers (CCSSO), Colorado Springs, CO.

*Bennett, R. E., & Ragosta, M. (1985). The technical characteristics of postsecondary admissions tests for handicapped examinees: A review of research. Journal of Special Education, 19, 255-267.

*Bennett, R. E., & Rock, D. A. (1989). Differential Item Function on the SAT-M Braille edition. Journal of Educational Measurement, 26, 67-79.

*Bennett, R. E., Rock, D. A., & Kirele, T. (1987b). GRE score level, test completion, and reliability for visually impaired, physically handicapped, and nonhandicapped groups. Journal of Special Education, 21, 9-21.

*Blaskey, P., Scheiman, M., Parisi, M., Ciner, E. B., Gallaway, M., & Selznick, R. (1990). The effectiveness of Irlen filters for improving reading performance: A pilot study. Journal of Learning Disabilities, 23(10), 604-612.

*Braun, H., Ragosta, M., & Kaplan, B. (1986). The predictive validity of the Scholastic Aptitude Test for disabled students (ETS Research Report 86-42). Princeton, NJ: Educational Testing Service.

Busk, P. L., & Marascuilo, L. A. (1987). Large-sample procedures for comparing and combing effect sizes within a single study. Journal of Experimental Education, 56, 8-14.

*Caudery, T. (1990). The validity of timed essay tests in the assessment of writing skills. ELT Journal, 44(2), 122-130.

Chiu, C. W. T. (1999). Scoring Performance Assessment Based on Judgements: Utilizing Meta-Analysis to Estimate Variance Components in Generalizability Theory for Unbalanced Situations. Unpublished Ph.D. dissertation. Michigan State University.

*Chen, H. C., & Graves, M. F. (1995). Effects of previewing and providing background knowledge on Taiwanese college students' comprehension of American short stories. TESOL Quarterly, 29(4), 663-686.

[1] A star preceding the reference indicates research studies meta-analyzed in the current report.

Crocker, L., & Algina, J. (1986). Introduction to Classical and Modern Test Theory. New York, NY: Rinehart and Winston.

*Gallina, N. B. (1989). Tourette's Syndrome Children: Significant Achievement and Social Behavior Variables (Tourette Syndrome, Attention Deficit Hyperactivity Disorder). Unpublished Ph.d. dissertation, City University of New York.

García, G. E., & Pearson, P. D. (1994). Assessment and diversity. In L.Darling-Hammond (Ed.), Review of Research in Education Vol. 20 (pp.337-392). Washington, DC: American Educational Research Association.

*Graves, A., Semmel, M., & Gerber, M. (1994). The effects of story prompts on the narrative production of students with and without learning disabilities. Learning Disability Quaterly, 17, 154-164.

*Halla, J. W. (1988). A Psychological Study of Psychometric Differences in Graduate Record Examinations General Test Scores Between Learning Disabled and Non-Learning Disabled Adults. Unpublished Ed.d., Texas Tech University.

Harwell, M. (1997). An investigation of the Raudenbush (1988) test for studying variance heterogeneity. Journal of experimental Education, 65(2), 181-90.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for Meta-Analysis. New York: Academic Press.

*Helwig, Rozek-Tedesco, Heath, Tindal, & Almond. (In Preparation). Reading as an access to math problems solving on multiple choice tests.

Heubert, J. P., & Hauser, R. M. (1998). Testing for Tracking, Promotion, and Graduation. Washington, D.C.: National Academy Press.

*Hill, G. A. (1984). Learning Disabled College Students: the Assessment of Academic Aptitude. Unpublished Ph.D. dissertation, Texas Tech University.

*Horton, S. V., Lovitt, T. C., & Bergerud, D. (1990). The effectiveness of graphic organizers for three classifications of secondard students in content area classess. Journal of Learning Disabilities, 23(1), 12-22.

*Jarvis, K. A. (1996). Leveling the Playing Field: a Comparison of Scores of College Students With and Without Learning Disabilities On Classroom Tests. Unpublished Ph.D. dissertation, The Florida State University.

*Jensen, H. K. (1997). Differences in Reading Comprehension Between College Students With Learning Disabilities and College Students Without Learning Disabilities On the Nelson-Denny Reading Test As Related to Question Type and Length of Test. Unpublished Ph.D. dissertation, The University of North Dakota.

Kalaian, H. A., & Becker, B. J. (1996, April). Modeling differences in variability. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Koretz, D. (1997). CRESST technical report on accommodations in Kentucky, CRESST.

*Linder, E. A. (1989). Learning Disabled College Students: a Psychological Assessment of Scholastic Aptitude. Unpublished Ed.d., Texas Tech University.

MacGlashan, S. (1999, May 2). New Rules to Include Special Ed Students. Times-Picayune, pp. BT, 1:6.

*MacArthur, C. A., & Haynes, J. B. (1995). Student assistant for learning from text (SALT): A hypermedia reading aid. Journal of learning disabilities, 28(3), 150-159.

McDonnell, L. M., McLaughlin, M. J., & Morison, P. (1997). Educating One & All: Students with Disabilities and Standards-Based Reform. N W , Washington DC: National Academy Press.

Mehrens, W. A., J. Millman, et al. (1994). Accommodations for candidates with disabilities. The Bar Examiner: 33-47.

*Miller, P. (1990). Use of the Peabody Picture Vocabulary Test-Revised (Ppvt- R) With Individuals With Severe Speech and Motor Impairment: Effect of Response Mode On Test Results (Speech Impairment). Unpublished Ph.D. dissertation, University of Kansas.

*Montani, T. O. (1995). Calculation Skills of Third-Grade Children With Mathematics and Reading Difficulties (Learning Disabilities). Unpublished Ed.d., Rutgers the State University of New Jersey - New Brunswick.

*Munger, G. F., & Loyd, B. H. (1991). Effect of speededness on test performance of handicapped and nonhandicapped examinees. Journal of Educational Research, 85(1), 53-57.

Neuburger, W. (1997). Oregon statewide assessment program. Paper presented at the National Large-Scale Assessment Conference organized by the Council of Chief State School Officers (CCSSO), Colorado Springs, CO.

*Ofiesh, N. S. (1997). Using Processing Speed Tests to Predict the Benefit of Extended Test Time For University Students With Learning Disabilities. Unpublished Ph.D. dissertation, The Pennsylvania State University.

Olson, J., F., & Goldstein, A. A. (1997). The Inclusion of Students with Disabilities and Limited English Proficient Students in Large-Scale Assessments (Research and Development Report NCES 97-482): Office of Educational Research and Improvement.

Pearson, P.D. (June, 1998). What do cognitive labs tell us about student knowledge? Paper presented at the National Large-Scale Assessment Conference organized by the Council of Chief State School Officers (CCSSO), Colorado Springs, CO.

Perlstein, L. (1999, April 4). New Tests Worry Special-Ed Advocates; Students May Lose. The Washington Post, pp. C, 1:1.

Phillips, S. E. (1994) High-Stakes Testing Accommodations: Validity versus Disabled Rights. Applied Measurement in Education 7(2): 93-120.

Porter, P. (1997). Accommodating students with disabilities in large-scale assessment programs: The Texas approach. Paper presented at the National Large-Scale Assessment Conference organized by the Council of Chief State School Officers (CCSSO), Colorado Springs, CO.

Raudenbush, S. W. (1988). Estimating change in dispersion. Journal of Educational Statistics, 13(2), p.148-171.

*Robinson, G. L. W., & Conway, R. N. F. (1990). The Effects of Irlen Colored Lenses on Students' Specific Reading Skills and Their Perception of Ability: A 12-Month Validity Study. Journal of Learning Disabilities, 23(10), 589-96.

*Ross, D. (1995). Effects of Time, Test Anxiety, Language Preference, and Cultural Involvement On Navajo Ppst Essay Scores. Unpublished Ed.d., Northern Arizona University.

*Runyan, M. K. (1991). The effect of extra time on reading comprehension scores for university students with and without learning disabilities. Journal of Learning Disabilities, 24(2), 104-107.

*Shepard, L. A., Taylor, G. A., Betebenner, D., & Weston, T. (In Press). Inclusion of Limited-English-Proficient students in Rhode Island's grade 4 mathematics performance assessment .

*Swain, C. R. (1997). A Comparison of a Computer-Administered Test and a Paper and Pencil Test Using Normally Achieving and Mathematically Disabled Young Children. Unpublished Ph.D. dissertation, University of North Texas.

Swanson, H. L. (1991). Operational Definitions and Learning Disabilities: An Overview. Learning Disability Quarterly, 14(4), 242-54.

*Tachibana, K. K. (1986). Standardized Testing Modifications For Learning Disabled College Students in Florida (Modality). Unpublished Ph.D. dissertation, University of Miami.

Tamar, L. (1998, Sep 20). Shaky Crutch for the Learning-Disabled. New York Times, pp. 4, 1:3.

Thurlow, M. L. (1997, March 24). Highlights of the accountability systems in two states that include all students with disabilities. Paper presented at the annual conference of the American Educational Research Association, Chicago, IL.

Thurlow, M., Ysseldyke J., & Olsen K. Self-Study Guide for the Development of Statewide Assessments that Include Students with Disabilities. [Online] Available http://www.coled.umn.edu nceo/OnlinePubs/self_study_guide.html, May 16, 1999.

Thurlow, M. L., Ysseldyke, J. E., & Silverstein, B. (1993). Testing Accommodations for Students with Disabilities: A review of the literature (Synthesis Report 4). National Center on Educational Outcome.

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., Harniss, M. (In Press). Accommoding students with disabilities on large-scale tests.

Trent, R. (1997). Ohio Department of Education Division of Assessment and Evaluation. Paper presented at the National Large-Scale Assessment Conference organized by the Council of Chief State School Officers (CCSSO), Colorado Springs, CO.

*Weaver, S. M. (1993). The Validity of the Use of Extended and Untimed Testing For Postsecondary Students With Learning Disabilities (Extended Testing). Unpublished Ed.d., University of Toronto (Canada).

Willingham, W. W., Ragosta, M., Bennett, R. E., Braun, H., Rock, D. A., & Powers, D. E. (1988). Testing Handicapped People: Allyn and Bacon, Inc.

Ysseldyke, J., Thurlow, M., Kozleski, E., & Reschly, D. (1998). Accountability for the Results of Educating Students with Disabilities. [Online] Available http://www.coled.umn.edu/nceo/OnlinePubs/awgfinal.html